線型回帰における
異常値の検出のための残差検定

原田桂一郎

目　　次

1　はじめに

線型回帰モデルの誤差項（攪乱項）には，正規性の仮定あるいは Causs-Markoff の仮定が設けられる。こうして相互に独立な観測値は線型回帰式によって規定される平均と一定分散をもつ正規分布，あるいは Gauss-Markoff の仮定の下では一定分布の母集団から生じたものとなる。観測値の中に，こうした分布とは異る別の分布をなす母集団から生じたものが存在するとき，そうした観測値を異常値（outliers）という。

最小二乗推定量は誤差項の正規性の下で，一様最小分散不偏推定量（uniformly minimum variance unbiased estimator）となる（それは最尤推定量でもある）。しかし異常値の存在による誤差項の正規性からの乖離は，最小二乗法の推定効率を損ね，そして回帰係数のｔ検定やＦ検定の結論を歪める。最小二乗推定量は Gauss-Markoff の仮定[①]の下でも最良線型不偏推定量（best linear unbiased estimator）となるが，異常値の存在によってこの仮定にも抵触すると，その特性は失なわれる。

回帰分析をすすめるうえで，観測値の中に異常値が存在することによって発生する現実的な問題は，回帰係数の最小二乗推定値，標準誤差，t 値などを歪めてしまうことである。実際の分析にあたって，異常値の存在を診断し，検出して，観測値の中の異常値を取り除いて推定を行うことが望ましい。

以下では，最小二乗法によって推定される回帰係数が，それぞれの観測値によってどれだけ影響を受けているかを測る二つの統計量から，影響の大きさは究極的には残差（residuals）にあらわれることを示す。さらに残差を診断し異常値を検出するための三つの統計的検定方式を提示する。二つの検定方式は誤差項に正規性の仮定を置いた場合に適用でき，一つはその分布を特定しない Gauss-Markoff の仮定の下でも適用できるものである。

注

① 誤差項に正規性の仮定が置けない場合には，回帰係数の推定結果に対し t 検定，F 検定は適用できない。しかし，t 検定は正規性からの乖離に対して頑健性（robustness）を有すことが証明されている。また，F 検定においては頑健性がないことが明らかとなっている。

検定方式の頑健性の問題については，たとえば Kendall and Stuart〔5〕, chapter 31. 参照。

2 異常値と残差の関係

次の標準線型回帰モデルを考える。

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{u}, \tag{2-1}$$

$\boldsymbol{y}$：被説明変数ベクトル $\boldsymbol{y}=(y_1\cdots\cdots y_N)'$,

$\boldsymbol{X}$：説明変数行列 $(N\times k)$,

$\boldsymbol{\beta}$：回帰係数ベクトル $\boldsymbol{\beta}=(\beta_1\cdots\cdots\beta_k)'$,

$\boldsymbol{u}$：誤差ベクトル $\boldsymbol{u}=(u_1\cdots\cdots u_N)'$，

$$\mathrm{E}(\boldsymbol{u})=\boldsymbol{o},\quad V(\boldsymbol{u})=\sigma^2 I\quad ,I(N\times N),$$

そして，(2-1) が推定されたものを，

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{b}+\boldsymbol{e}, \tag{2-2}$$

$\boldsymbol{y}$：(2-1) に同じ，

$\boldsymbol{X}$：(2-1) に同じ，

$\boldsymbol{b}$：βの推定量，

$\boldsymbol{e}$：残差ベクトル $(N\times 1)$，

S^2：誤差分散 σ^2 の推定量，

と表わす。さらに次の表記を与えておく。

$\boldsymbol{x}_i$　：$\boldsymbol{X}$行列の第 i 行，

$\boldsymbol{X}(i)$：$\boldsymbol{X}$行列から第 i 行を除いたもの，

$\boldsymbol{y}(i)$：$\boldsymbol{y}$ベクトルから第 i 要素を除いたもの，

$\boldsymbol{b}(i)$：$\boldsymbol{X}$と$\boldsymbol{y}$から第 i 行，第 i 要素を除いた場合の β の推定量，

$S^2(i)$：$\boldsymbol{X}$と$\boldsymbol{y}$から第 i 行，第 i 要素を除いた場合の σ^2 の推定量。

それぞれの変数の各観測値は行列$\boldsymbol{X}$とベクトル$\boldsymbol{y}$の行と要素をなしている。影響力の強い観測値は回帰係数，標準誤差，t 値などの推定値に大きなインクパトを与える。各観測値が様々な推定値に与える影響を調べる一つの方法は行列$\boldsymbol{X}$とベクトル$\boldsymbol{y}$の各行を除いて計算した推定量の値を調べることである。ある行を除外して計算した推定値に大きな変化があれば，その行における観測値が推定値に与えた影響が大きいと判定できる。

ここで，回帰係数推定値に各観測値が与える影響を調べるために，回帰係数の最小二乗推定量$\boldsymbol{b}$と（$\boldsymbol{X}$，$\boldsymbol{y}$）の任意の行を除いた推定量$\boldsymbol{b}(i)$の差$\Delta\boldsymbol{b}(i)$を考えてみる。Belsley, Kuh and Welsch〔2〕はこの統計量を次のように与えている。②

$$\Delta b(i)=b-b(i)=\frac{(X'X)^{-1}x_i'e_i}{1-x_i(X'X)^{-1}x'_i} \tag{2-3}$$

$$e_i=y_i-x_i b,\ \ i=1,\cdots\cdots,N,$$

なお，$1-x_i(X'X)^{-1}x_i'$ は行列 $M=I-X(X'X)^{-1}X'$ の i 番目の対角要素である。

この $\Delta b(i)$，$i=1,\cdots\cdots N$，の系列の変動が大きい場合，それが観測値のコーディングに誤りがあったか，モデルに適合するに足る観測値がなかったか，あるいはモデル特定化の誤りがあったことに対する警告である。モデルやデータにこのような事態がなく，特定の $\Delta b(i)$ の値が大きいことが見い出せると，i 番目の観測値が係数推定値に大きなインパクトを与えているといえよう。$\Delta b(i)$ 系列は各観測値が回帰係数の推定値に与える影響度を示している。

次に回帰係数の最小二乗推定値に大きな影響を与える観測値を検出する，より直接的な方法を示す。まず，i 番目の観測値にウエイトを付ける。回帰係数の最小二乗推定量をそのウエイトについて微分し，そのウエイトの値を1とした統計量を求める。この統計量が i 番目の観測値のウエイトの僅かな変化に対する回帰係数の最小二乗推定値の感度を測ることになる。

i 番目の観測値にウエイトを付けるための行列を次のように定めておく。

$$V=\begin{bmatrix} 1 & 0 & \cdots\cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \sqrt{w_i} & \\ & & & 1 \\ 0 & \cdots\cdots & & 1 \end{bmatrix},\quad w_i\geq 1$$

この行列を（2—1）の両辺に乗じた

$$Vy=VX\beta+Vu, \tag{2-4}$$

は正規方程式，

$\boldsymbol{X}'V'V\boldsymbol{y}=\boldsymbol{X}'V'V\boldsymbol{X}\boldsymbol{b}(w_i)$,

$\boldsymbol{b}(w_i)$：観測値にウエイトを付けたときの（2-4）の $\boldsymbol{\beta}$ の最小二乗推定量，

を与えるから，

$$\boldsymbol{b}(w_i)=(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}, \tag{2-5}$$

$$W=V'V=\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & w_i & \\ & & & \ddots \\ 0 & \cdots & & 1 \end{bmatrix},$$

が得られる。$\boldsymbol{b}(w_i)$ を w_i について微分したものを Belsley *et al.*［2］③は次のように導出している。

$$\frac{\partial \boldsymbol{b}(w_i)}{\partial w_i}=\frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i e_i}{[1-(1-w_i)\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i]^2} \tag{2-6}$$

また，$w_i=1$ としたものは

$$\boldsymbol{\Phi b}(w_i)=\frac{\partial \boldsymbol{b}(w_i)}{\partial w_i}\Big|_{w_i=1}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i e_i,\quad i=1,\cdots,N. \tag{2-7}$$

$\boldsymbol{\Phi b}(w_i)$，$i=1,\cdots,N.$ の系列を調べて i 番目の値が大である場合，i 観測値が最小二乗推定による回帰係数 $\boldsymbol{b}$ に大なる影響を与えていると判定できる。

ところで，$\boldsymbol{\Delta b}(i)$ と $\boldsymbol{\Phi b}(w_i)$ において，主要な要素は，$\boldsymbol{X}$ 行列の観測値の情報を与える $\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i$ と $\boldsymbol{y}$ ベクトルの観測値の情報を与える残差 e_i である。$0\leqq \boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i \leqq 1$ であるからその変動は小さいが④，残差は回帰係数の最小二乗推定値に大きな影響を及ぼす観測値（異常値）の検出の指標となる。異常値は残差，$e_i=y_i-\hat{y}_i$，の値を特に大きくする。残差平方和最小の規準に基づく最小二乗推定では，異常値による残差が大

きなウエイトを占め回帰係数推定値を歪めてしまう。この傾向は標本の大きさ（sample size）があまり大きくない場合に顕著である。

残差は回帰分析の結果に重大な影響を及ぼす事態を検出するために既に使用されている。すなわち，誤差項の分散の不均一性（heteroscedasticity）や自己相関の検出，あるいは誤差項の正規性の検証などに於て活用されている。

注

② (2-3) の証明は付1参照。

③ (2-6) と (2-7) の導出は付2に示す。

④ $e_i = y_i - \boldsymbol{x}_i \boldsymbol{b}$

$= y_i - \boldsymbol{x}_i (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{X}'\boldsymbol{y}$

$= y_i - \boldsymbol{x}_i (\boldsymbol{X}'\boldsymbol{X})^{-1} (\boldsymbol{X}'(i)\boldsymbol{y}(i) + \boldsymbol{x}'_i y_i)$

$= (1 - \boldsymbol{x}_i (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{x}'_i)\ y_i - \boldsymbol{x}_i (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{X}'(i)\ \boldsymbol{y}\ (i)$,

$e_i = O$, のときに $\boldsymbol{x}_i (\boldsymbol{X}'\boldsymbol{X})\ \boldsymbol{x}'_i = 1$ となる。また，異常値 $\boldsymbol{x}_i$ によって e_i が大となる場合，$\boldsymbol{x}_i (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{x}'_i$ は零に近づく。

3 残差の検定

異常値の検出は $\Delta \boldsymbol{b}(i)$ や $\Phi \boldsymbol{b}(w_i)$ の系列を観測することによって可能であるが，それは両統計量の共通要素である残差が異常値によってその値が大となることによるのであるから，残差系列を診断することに帰着する。しかし，値の大なる残差を異常値によるものと判定するには，その統計的検定を行うことが合理的である。

異常値は一定分布の母集団とは異なる別の分布の母集団から発生した観測値である。異常値の検出は，この観点に基づいて残差の検定を通して間接的に行う。一定分布にしたがうという前提の下での残差の中に，別個の異なる分布にしたがうものが存在するとき，その残差を構成する観測値は，平均 $\boldsymbol{X}\boldsymbol{\beta}$，分散 $\sigma^2 \boldsymbol{I}$ の分布をなす母集団とは異なる別の母集団から抽出された異常値であると判定する。

最初に提示する方式は，残差の各々について検定を行うものである。いま，回帰式の誤差項ベクトルが，$N(\boldsymbol{O},\ \sigma^2\boldsymbol{I})$ にしたがうと仮定される場合，各々の残差は

$$e_i \sim N\left[O,\ \sigma^2(1-\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i)\right] \tag{3-1}$$

である⑤。ここで検定すべき帰無仮説は，$Ho: E(e_i)=O$，である。この帰無仮説の下で e_i を標準化したものは，

$$Z=\frac{e_i}{\sqrt{\sigma^2(1-\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i)}} \sim N\ (0,\ 1), \tag{3-2}$$

となる。また，正規母集団，$\boldsymbol{u}\sim N\ (\boldsymbol{O},\ \sigma^2\boldsymbol{I})$，から e_i を選んで

$$U^2=\frac{\sum_{i=1}^{N} e_i^2}{\sigma^2}=\frac{(N-k)S^2}{\sigma^2}, \tag{3-3}$$

$$S^2=\frac{\sum_{i=1}^{N} e_i^2}{N-k},$$

をつくると，この量は自由度 $N-k$ の χ^2 分布にしたがう。よって，

$$e_{st}=Z/\sqrt{U^2/N-k}=\frac{e_i}{S\sqrt{1-\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i}} \tag{3-4}$$

は自由度 $N-k$ の t 分布にしたがう。

残差は正と負の値をとるから，帰無仮説 $E(e_i)=O$ の下で e_{st} が一定の範囲に入る確率は，$t_{\alpha/2}$ を e_{st} の確率 $\alpha/2$ の点とすれば，

$$P\gamma\ [-t_{\alpha/2}\leqq e_{st}\leqq t_{\alpha/2}]=1-\alpha,\ \ \alpha:\text{有意水準}$$

である。したがって，e_{st} が

$$|e_{st}|\geqq t_{\alpha/2} \tag{3-5}$$

となった場合，帰無仮説は棄却される。その残差は (3-1) に示されるもの

とは別の分布にしたがうものであり，それを構成する観測値は異常値と判定される。したがって，(3-4) と (3-5) によれば異常値のポジションが推定できる。

なお，Belsley *et al.*〔2〕は標本の大きさに影響されない統計量であるとして，

$$e_{st}^{*}\equiv\frac{e_i}{S(i)\sqrt{1-\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i}} \tag{3-6}$$

を提示して（e_{st}^{*}は自由度 $N-k-1$ の χ^2 分布をなす），確率$\alpha/2$点，$t_{\alpha/2}$，を2と定めている⑥。

以下に示す二つの検定方式は，残差 e_i $(i=1,\cdots,N)$ をその値の大きさからみて異常値によって構成されていると考えられるものと，そうでないものとに二分し，分布の同質性を検定するものである。

残差（N個）を，その値が大であり異常値によって構成されたと考えられるm個を一つのグループとし，またその値が小であるものn個を他のグループとして二つに分離する。最初のグループに属す残差は分布Fにしたがい，他は分布Gにしたがい，相互に独立であると仮定する。ここでの帰無仮説はF＝Gであり，対立仮説はF≠Gである。帰無仮説が棄却されたならば，両者の分布は異なり，最初のグループの残差の各々を構成している観測値は異常値と判定される。

いま，誤差項の正規性を仮定すると，t検定を適用することができる。二分した残差，e_{11}，$e_{12},\cdots$，e_{1m}，と e_{21}，e_{22}，$\cdots e_{2n}$，が共通の未知の分散σ_e^2で，未知の平均 $E(e_{1i})=\mu_1$，$E(e_{2l})=\mu_2$ をもって独立に正規分布をなすものと仮定する⑦。この場合の帰無仮説は $F=G$ であるから，$H_0:\mu_1=\mu_2$，対立仮説は $H_a:\mu_1\neq\mu_2$ である。ここで，

$$\bar{e}_1=\frac{1}{m}\sum_{j=1}^{m}e_{1j},\qquad \bar{e}_2=\frac{1}{n}\sum_{l=1}^{n}e_{2l},$$

とすると，$\bar{e}_1$ と $\bar{e}_2$ の分布は

$$\bar{e}_1 \sim N(\mu_1,\ \sigma_e^2/m), \tag{3-7}$$

$$\bar{e}_2 \sim N(\mu_2,\ \sigma_e^2/n), \tag{3-8}$$

である。$\bar{e}_1$ と $\bar{e}_2$ の差は

$$(\bar{e}_1-\bar{e}_2) \sim N\left[(\mu_1-\mu_2),\ \sigma_e^2\left(\frac{1}{m}+\frac{1}{n}\right)\right] \tag{3-9}$$

となる。仮説 H_0 の下では $\mu_1-\mu_2=0$ であるから，

$$Z=\frac{\bar{e}_1-\bar{e}_2}{\sigma_e\sqrt{\frac{1}{m}+\frac{1}{n}}} \sim N\ (0,\ 1) \tag{3-10}$$

となる。他方，

$$\sum_{j=1}^{m}(e_{1j}-\bar{e}_1)^2/\sigma_e^2,$$

$$\sum_{l=1}^{n}(e_{2l}-\bar{e}_2)^2/\sigma_e^2,$$

はそれぞれ（$m-1$），（$n-1$）の自由度をもって独立に χ^2 分布をなすから，両者の和

$$V \equiv \frac{\sum(e_{1j}-\bar{e}_1)^2+\sum(e_{2l}-\bar{e}_2)^2}{\sigma_e^2} \tag{3-11}$$

は自由度（$m+n-2$）の χ^2 分布となる。

したがって，

$$t=\frac{Z}{\sqrt{V/(m+n-2)}}=\frac{\bar{e}_1-\bar{e}_2}{S\sqrt{\frac{1}{m}+\frac{1}{n}}}, \tag{3-12}$$

$$S=\sqrt{\left\{\sum_{j=1}^{m}(e_{1j}-\bar{e}_1)^2+\sum_{l=1}^{n}(e_{2l}-\bar{e}_2)^2\right\}\Big/(m+n-2)},$$

は自由度（$m+n-2$）の t 分布となる。帰無仮説（$\mu_1=\mu_2$）の下で t が一定値以下となる確率は，t_α を t の確率 α の点とすれば，

$$P\gamma\,[\,t \leq t_a\,] = 1-\alpha,\quad \alpha：有意水準$$

であるから，

$$t \geqq t_a \tag{3-13}$$

の範囲に t が入れば，仮説 $H_0：\mu_1=\mu_2$ が棄却され，e_{1j} と e_{2l} は異なる別個の分布から発生すると判定される。

次に，回帰式の誤差分布を特定せず，Gauss-Markoff 仮定を置いた場合の検定方式は順位検定が適切である。順位検定は最初に Wilcoxon〔14〕によって提示され，Mann and Whitney〔7〕によって展開された方式で，Wilcoxon 検定（または Mann-Whitney 検定）と呼ばれている。これは，分布によらない検定もしくはノンパラメトリック検定と呼ばれる検定方式の領域に入るものである。

ここで，残差の系列 e_i，$i=1,\cdots,N$，にその値の小なる順に順位（rank）を付ける。異常値が定める残差はその値が大となるので，順位は後位になる。異常値によるものと考えられる順位が後位（最後からm個）の残差を一つのグループとし，順位が1からnまでの残差を他のグループとする。e_{11}，e_{12}，$\cdots$，e_{1n} と e_{21}，e_{22}，$\cdots e_{2m}$ がそれぞれ値の小さなものから順に配列される。

前のグループの母集団分布をF，後のグループの母集団分布をGとする。順位が明確に定められるのは，e_{1p}，$P=1,\cdots n$，と e_{2s}，$S=1,\cdots m$，の間に同じ値（tie）がないときだけである。したがって，上記の配列は，tie がおこらない，$Pr(e_{1p}=e_{2s})=0$，ことが確率1の唯一の事象であると仮定する。そのための十分条件は，分布(関数)FとGの連続性である。

また，値U（これからの検定に用いる統計量である）は，e_{2s} グループの各々の値が，e_{1p} グループの各々の値を超えている回数とする。

検定しようとする帰無仮説は，H_0: F＝Gであり，対立仮説は H_a: F≠Gである。帰無仮説の下で，Uの値が十分に小さくなる確率，$P\gamma(U\leq\bar{U})$，が有意水準よりも小さく，$P\gamma(U\leq\bar{U}) \leq \alpha$ であれば，帰無仮説は棄却さ

れる。

Mann and Whitney〔7〕は統計量Uとその分布を次のように与えている。

$$U=m\,n+\frac{m(m+1)}{2}-T \tag{3-14}$$

T：e_{2s} の各々の順位の和

いま，e_{2s} の各々の値が e_{1p} の各々の値をU個超えるような e_{1p} と e_{2s} の配列の数を $\bar{P}_{nm}(U)$ で表わす。また帰無仮説の下で，全体（$n+m=N$個）を合併して，e_{2s} グループの順位としてm個を割り当てることができる ${}_{n+m}C_m$ 個の e_{1p}，e_{2s} の並べ方が，いずれも確率 $1/{}_{n+m}C_m$ をもつ。したがって，e_{2s} グループの各々が e_{1p} の各々をU個超える配列が生ずる確率は，

$$P_{nm}(U)=\frac{\bar{P}_{nm}(U)}{{}_{n+m}C_m} \tag{3-15}$$

である。

さらに，$n\geqq 3$，$m\geqq 1$ から $n\leqq 8$，$m\leqq 8$，まで，$P_{nm}(U)$を計算して得た確率が表としてまとめられている（Mann and Whitney〔14〕，*pp.* 52—54.）。なお，$n=m=8$ については，Uの確率は正規分布の値と極く僅かな差しかないことが，その表から読み取れる。この事はnとmが大きくなるにしたがって，Uの分布は正規分布に近づくことを示唆する。

Mann and Whitney〔14〕はUの漸近分布が正規分布となり，平均は，

$$\mu_u=nm/2, \tag{3-16}$$

であり，これはUの平均に等しく，また分散は，

$$\sigma_u{}^2=nm(n+m+1)/12 \tag{3-17}$$

これはUの分散に等しいことを証明している。

したがって，n，mが大きい場合，

$$\boldsymbol{n}=\frac{U-\mu_u}{\sqrt{\sigma_u^2}}\sim N\ (0,\ 1) \tag{3-18}$$

であるから，$\boldsymbol{n}$が一定値以下になる確率は，n_α を$\boldsymbol{n}$の確率αの点とすれば，

$$P\gamma[\boldsymbol{n}\leq n_\alpha]=1-\alpha,\ \alpha：\text{有意水準}$$

となる。$\boldsymbol{n}$の値が

$$\boldsymbol{n}\geq n_\alpha \tag{3-19}$$

の範囲に入れば，仮説 H_0 は棄却される。

ところで，誤差項の分布に正規性を仮定しない場合，正規性を前提とした検定が適用できないので特定の分布を前提としない Wilcoxon 検定が適切であるとした。しかし，検定方式で重要なポイントは検出力，すなわち仮説が正しくないときに仮説を棄却する確率である。ここで，検出力に基づいて両検定方式を比較して，Wilcoxon 検定の性能を確認しておく。

両者の検定の大きさ（棄却域の大きさ）が等しくαであるとし，同じ対立仮説に対し等しい検出力βを持つようにするとき，検定に必要な観測値の数を，t 検定では$m^*=n^*$，Wilcoxon 検定では$m=n$とし，その比n^*/m（これを Wilcoxon 検定の t 検定に関する効率という）を考える。nが無限に大きくなるにしたがって，比 n^*/m がαとβに無関係な極限値εに接近すると仮定する。この ε（効率の極限値）を t 検定に対する Wilcoxon 検定の漸近的効率（asymptotic efficiency）と呼ぶ。これを用いれば，等しい検出力を得るのに Wilcoxon 検定が t 検定に比べてどのくらいの標本の大きさを要するかが解る。[8]

分布FとGが同じ分散をもつ正規分布であるとき，$\varepsilon=3/\pi\fallingdotseq 0.95$と与えられている（Hodges and Lehmann〔4〕, p. 325.）。したがって，正規分布を前提とすると，t 検定の方が優れているが，この場合 Wilcoxon 検

定による効率の損失は約5パーセントである。また，有限分散をもつすべての分布に対して，$\varepsilon \geq 0.864$であることが示され（Hodges and Lehmann〔4〕, p. 325.），分布が特定できない場合，Wilcoxon 検定は優れた検定方式といえる。

注

⑤ (3-1) は付3参照。

⑥ Belsley *et al.*〔1〕, p. 28. 両側検定で$\alpha = 5\%$とすると $t_{\alpha/2}$ の値は約2となる。

⑦ 残差の各々を，その値の大きさによって二のグループに分けたので，それぞれのグループの残差は，(3-1) とは別の未知の平均，共通の分散の正規分布FとGにしたがうと仮定する。

⑧ たとえば，$\varepsilon = 1/2$ である場合，等しい検出力を得るのに Wilcoxon 検定は t 検定の2倍の標本を必要とする。

4 むすび

本稿では，線型回帰モデルの回帰係数の最小二乗推定値が，それぞれの観測値によって与えられる影響度を，統計量 $\Delta \boldsymbol{b}(i)$ と $\Phi \boldsymbol{b}(w_i)$ によって診断できることを示した。これらの統計量の中で，残差が重要な要素となっているので，残差の診断によって異常値を存在を知ることができる。

回帰係数のそれぞれの観測値に対する感度（安定度）を診断するうえで，$\Delta \boldsymbol{b}(i)$ と $\Phi \boldsymbol{b}(w_i)$ が重要な情報を伝えるので，Storer and Crowley [12] は Cox Regression Model についての $\Delta \boldsymbol{b}(i)$ を，また Polasek [9] は General Linear Model についての $\Phi \boldsymbol{b}(w_i)$ を導出している。

異常値による残差と判定するための検定方式として，誤差項に正規性を仮定した場合には t 検定が適用できる。しかし標準線型回帰モデルの場合，その検定方式は Wilcoxon 検定が適切である。誤差項に正規性を仮定できないので，特定の分布に依拠しないノンパラメトリックな検定方式を

採ること，それよりも重要な理由は Wilcoxon 検定の検出力が優れているからである。正規性の前提の下では t 検定の検出力の方が強力であるが，この場合に Wilcoxon 検定による効率の損失は 5 パーセントにすぎない。また，正規性が保たれない様々な分布の場合，Wilcoxon 検定は効率のうえでかなり有利であることが明らかにされている。

付録　結 果 の 証 明

付 1　(2-3) の証明

$$\boldsymbol{X}'\boldsymbol{X}=\boldsymbol{X}'(i)\boldsymbol{X}(i)+\boldsymbol{x}'_i\boldsymbol{x}_i$$

であるから，

$$\begin{aligned}(\boldsymbol{X}'(i)\boldsymbol{X}(i))^{-1}&=(\boldsymbol{X}'\boldsymbol{X}-\boldsymbol{x}'_i\boldsymbol{x}_i)^{-1}\\&=(\boldsymbol{X}'\boldsymbol{X})^{-1}\left[\boldsymbol{I}-\boldsymbol{x}'_i\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]^{-1}\\&=(\boldsymbol{X}'\boldsymbol{X})^{-1}\left[\boldsymbol{I}+\boldsymbol{x}'_i\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\right.\\&\quad\left.+\boldsymbol{x}'_i\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i\boldsymbol{x}(\boldsymbol{X}'\boldsymbol{X})^{-1}+\cdots\cdots\right]\\&=(\boldsymbol{X}'\boldsymbol{X})^{-1}+(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\\&\quad+(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i'\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}+\cdots\cdots\\&=(\boldsymbol{X}'\boldsymbol{X})^{-1}+(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\\&\quad\times\left[\boldsymbol{I}+\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i+\cdots\cdots\right]\\&=(\boldsymbol{X}'\boldsymbol{X})^{-1}+\frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}}{1-\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i},\qquad\text{(A-1)}\end{aligned}$$

よって，

$$(\boldsymbol{X}'(i)\boldsymbol{X}(i))^{-1}=(\boldsymbol{X}'\boldsymbol{X})^{-1}+\frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}}{1-\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i}\qquad\text{(A-2)}$$

また，(A-2) により，

$$\begin{aligned}\boldsymbol{b}-\boldsymbol{b}(i)&=\boldsymbol{b}-(\boldsymbol{X}'(i)\boldsymbol{X}(i))^{-1}\boldsymbol{X}'(i)\boldsymbol{y}(i)\\&=\boldsymbol{b}-\left\{(\boldsymbol{X}'\boldsymbol{X})^{-1}+\frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}}{1-\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i'}\right\}\boldsymbol{X}'(i)\boldsymbol{y}(i)\end{aligned}$$

であるから，ここで

$$\boldsymbol{X}'(i)\boldsymbol{y}(i)=\boldsymbol{X}'\boldsymbol{y}-\boldsymbol{x}'_i y_i,$$

$$\boldsymbol{b}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y},$$

$$e_i=y_i-\boldsymbol{x}_i\boldsymbol{b}$$

の関係を用いて右辺を整理すると

$$\boldsymbol{b}-\boldsymbol{b}(i)=\frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i\,e_i}{1-\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i} \tag{A-3}$$

付2　(2-6), (2-7) の証明

すでに

$$W=\begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}\quad \begin{matrix} 1 \\ 1 \\ w_i \\ 1 \end{matrix}\quad ,\ w_i\geq 1,$$

$$\boldsymbol{b}(w_i)=(\boldsymbol{X}'\,W\boldsymbol{X})^{-1}\boldsymbol{X}'\,W\boldsymbol{y} \tag{A-4}$$

が与えられている。

また,

$$\begin{aligned}\boldsymbol{X}'\,W\boldsymbol{X}&=\boldsymbol{X}'\boldsymbol{X}-\boldsymbol{x}'_i\,\boldsymbol{x}_i+w_i\,\boldsymbol{x}'_i\,\boldsymbol{x}_i\\&=\boldsymbol{X}'\boldsymbol{X}-(1-w_i)\,\boldsymbol{x}'_i\,\boldsymbol{x}_i\end{aligned}$$

であるから,

$$(\boldsymbol{X}'\,W\boldsymbol{X})^{-1}=[\boldsymbol{X}'\boldsymbol{X}-(1-w_i)\,\boldsymbol{x}'_i\,\boldsymbol{x}_i]^{-1},$$

(A-1) を用いると,

$$(\boldsymbol{X}'W\boldsymbol{X})^{-1}=(\boldsymbol{X}'\boldsymbol{X})^{-1}+\frac{(1-w_i)\,(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i\,\boldsymbol{x}_i\,(\boldsymbol{X}'\boldsymbol{X})^{-1}}{1-(1-w_i)\,\boldsymbol{x}_i\,(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i} \tag{A-5}$$

を得る。また,

$$\boldsymbol{X}'W\boldsymbol{y}=\boldsymbol{X}'\boldsymbol{y}-(1-w_i)\,\boldsymbol{x}'_i\,y_i \tag{A-6}$$

であるから, (A-4) に (A-5) および (A-6) を代入して

$$\begin{aligned}\boldsymbol{b}(w_i)=&\left\{(\boldsymbol{X}'\boldsymbol{X})^{-1}+\frac{(1-w_i)\,(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i\,\boldsymbol{x}_i\,(\boldsymbol{X}'\boldsymbol{X})^{-1}}{1-(1-w_i)\,\boldsymbol{x}_i\,(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i}\right\}\\&\times\{\boldsymbol{X}'\boldsymbol{y}-(1-w_i)\,\boldsymbol{x}'_i\,y_i\}\end{aligned}$$

$$= \boldsymbol{b} - (1-w_i)(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i\, y_i$$
$$+\frac{(1-w_i)(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}}{1-(1-w_i)\,\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i}\{\boldsymbol{X}'\boldsymbol{y}-(1-w_i)\boldsymbol{X}'\,y_i\}$$
$$= \boldsymbol{b} - \frac{(1-w_i)(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i(y_i-\boldsymbol{x}_i\boldsymbol{b})}{1-(1-w_i)\,\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i},$$

したがって，

$$\boldsymbol{b}(w_i) = \boldsymbol{b} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i\, e_i\left\{\frac{(1-w_i)}{1-(1-w_i)\,\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i}\right\} \quad (A\text{-}7)$$

が得られる。かくして，

$$\frac{\partial\,\boldsymbol{b}(w_i)}{\partial\, w_i} = \frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i\, e_i}{[1-(1-w_i)\,\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i]^2} \quad (A\text{-}8)$$

また $w_i = 1$ とおくと，

$$\left.\frac{\partial\,\boldsymbol{b}(w_i)}{\partial\, w_i}\right|_{w_i=1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_i\, e_i \quad (A\text{-}9)$$

付3　残差の分布

残差ベクトル $\boldsymbol{e}$ は次のように展開できる。

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}$$
$$= \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = M\boldsymbol{y}, \quad (A\text{-}10)$$
$$M = I - X(X'X)^{-1}X' \quad (A\text{-}11)$$
$$M'M = M,\ M：\text{n 次 Symmetric, idempotent matrix}$$

さらに，計算をすすめると

$$\boldsymbol{e} = \mathrm{M}\boldsymbol{y}$$
$$= M(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{u}), \quad (A\text{-}12)$$

ここで，

$$M\boldsymbol{X} = (\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\,\boldsymbol{X} = \boldsymbol{O}$$

であるから，

$$\boldsymbol{e} = M\boldsymbol{u}, \quad (A\text{-}13)$$
$$\boldsymbol{u} \sim N(\boldsymbol{O},\ \sigma^2\boldsymbol{I}). \quad (A\text{-}14)$$

Anderson［1］, p. 19, Theorem 2. 4. 1. により

$$\boldsymbol{e} \sim N(\boldsymbol{O},\ M\sigma^2\boldsymbol{I}M'),$$

$$M\sigma^2 \boldsymbol{I} M' = \sigma^2 M,$$

$$\boldsymbol{e} \sim N(\boldsymbol{O},\ \sigma^2 M) \qquad \text{(A-15)}$$

M の i 対角要素 $= 1 - \boldsymbol{x}_i (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{x}'_i$

$$\epsilon_i \sim N[O,\ \sigma^2 (1 - \boldsymbol{x}_i (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{x}'_i)] \qquad \text{(A-16)}$$

（1985年9月）

REFERENCES

〔1〕 Anderson, T. W. : *An Introduction to Multivariate Statistical Analysis.* New York : John Wiley & Sons, 1958.

〔2〕 Belsley, D. A., E. Kuh and R. E. Welsch : *Regression Diagnostics ; Identifying Influential Data and Sources of Collinearity.* New York : John Wiley & Sons, 1980.

〔3〕 Chernoff, H. and I. R. Savage : "Asymptotic Normarity and Efficiency of Certain Nonparametric Test Statistics," *Annals of Mathematical Statistics*, 29 (1958), 972-994.

〔4〕 Hodges, J. L., Jr. and E. L. Lehmann : "The Efficiency of Some Nonparametric Competitors of the t-Test, "*Annals of Mathematical Statistics*, 27 (1956), 324-335.

〔5〕 Kendall, M. and A. Stuart : *The Advanced Theory of Statistics*, Vol. 3, 4 th ed.. London : Charles Griffin, 1979.

〔6〕 Lehmann, E. L. : *Testing Statistical Hypotheses.* New York : John Wiley & Sons, 1959.

〔7〕 Mann, H. B. and D. R. Whitney : "On a Test of One of Two Random Variables is Stochastically Larger than the Other," *Annals of Mathematical Statistics*, 18 (1947), 50-60.

〔8〕 Mood, A. M. and F. A. Graybill : *Introduction to the Theory of Statistics.* New York : McGrow-Hill, 1963.

〔9〕 Polasek, W. : "Regression Diagnostics for General Linear Regression Models, "*Journal of the American Statistlcal Association*, 79 (1984), 336-340.

〔10〕 Rao, C. R. : *Linear Statistical Inference and Its Applications*, 2 nd ed.. New York : John Wiley & Sons, 1973.

〔11〕 Schweder. T. : "Some "Optimal" Methods to Detect Stractural Shift or Outliers in Regression, "*Journal of the American Statistical Association*, 71 (1976), 491-501.

〔12〕 Storer, B. E. and J. Crowley : "A Diagnostic for Cox Regression and General Conditional Likeliroods, "*Journal of the American Statistical Association*, 80 (1985), 139-147.

〔13〕 Theil, H. : *Principles of Econometrics*. Amsterdam : North Holland, 1979.

〔14〕 Wilcoxon, F. : "Individual Comparisosns by Ranking Method," *Biometrics Bulletin*, 1 (1945), 80-83.